Data Wrangling on Crawled Data

Mr. S Gokula Krishnan[1], Ms.K.Bhavitha[2], Ms.P.Sudha Rani[2]

[1]Assistant Professor, SCSVMV University
[2]Student, SCSVMV University

---***---

ABSTRACT

Data wrangling is the process of cleaning, structuring and enriching raw data into a desired format for better decision making in less time. It is mainly used to improve Data Quality. A crawler is a program that visits Websites and reads their pages and other information in order to create entries for a search engine index with the help of cleaning, structuring and unifying cluttered and complex data into sets, data wrangling ensures that data becomes easy to access and analyze. It makes certain that there is no unarranged stack of data during analysis.

Keywords- Crawled Data; Data Quality; Data Cleaning; Structuring;

INTRODUCTION

These days, data is what governs our everyday lives as well as business fortunes. They can come from diverse sources, at different times and are available in different formats. Within this data lies invaluable insights waiting to be gleaned by data scientists, but before that they would need the data in proper order and in a consistent format to be able to carry out analysis.

This can be processed by Data Verification, Data Validation and Data Enrichment.

METHODOLOGIES

1. VERIFICATION

Data Verification is a process in which different types of data are checked for accuracy and inconsistencies after data migration is done. It helps to determine whether data was accurately translated when data is transferred from one source to another, is complete and supports processes in the new system. During verification, there may be a need for a parallel run of both systems to identify areas of disparity and forestall erroneous data loss. A type of data verification is double data entry and proofreading data. Proofreading data involves someone checking the data entered against the original document. This is also time consuming and costly.

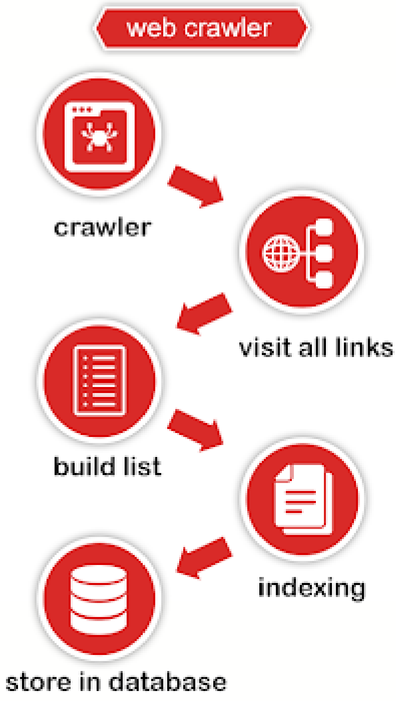
2. VALIDATION

Data Validation is the process of ensuring data have undergone data cleansing to ensure they have data quality, that is, that they are both correct and useful. It uses routines, often called "validation rules", "validation constraints", or "check routines", that check for correctness, meaningfulness, and security of data that are input to the system. The rules may be implemented through the automated facilities of a data dictionary, or by the inclusion of explicit application program validation logic of the computer and its application.

3. ENRICHMENT

After cleaning, it will have to be enriched .This means that you will have to take stock of what is in the data and strategies whether you will have to augment it using some additional data in order to make it better. You should also brainstorm about whether you can derive any new data from the existing clean data set that you have.

Data Wrangling is mainly used to clean the raw or crawled data which is taken from the web. The following image shows how the data is taken and validated and sent to the database.



A web crawler is a software program that visits websites and reads their pages and other related information in order to build entries for a search engine index. The major search engines like Google, Yahoo ,Bing etc on the web all have such a program, which is also known as a "web spider".

Minimized Data Leakage

Data Leakage is often considered one of the biggest challenges of Machine Learning. And since ML algorithms are used for data processing the threat grows exponentially. The thing is prediction relies on the accurateness of data. And if the calculated prediction is based on uncertain data this prediction is as good as a wild guess estimation.

WORKING MODEL

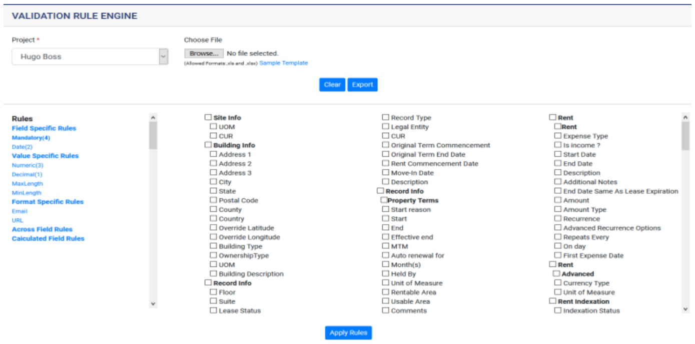
The working model contains Add, Edit and Delete function to send details into the database from the screen vice versa.

Validation Rules:

Validation rules verify that the data a user enters in a record meets the standards you specify before the user can save the record. A validation rule can contain a formula or expression that evaluates the data in one or more fields and returns a value of "True" or "False".

The projects which are need to be validated are entered into the database from those list of projects one by one is selected and processed according to the rules which are entered before.

The below screen is used to validate the project by using project attributes



VALIDATION RULE ENGINE

Project * Hugo Boss

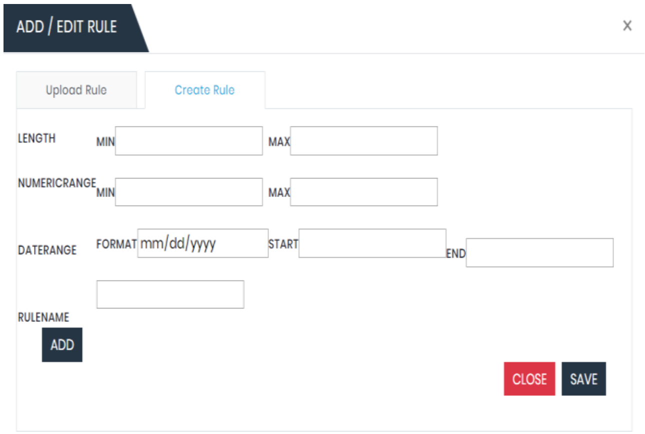
Choose File Browse... No file selected.
(Allowed Formats:.xls and .xlsx) Sample Template

A new project is uploaded below those projects were taken from the projects which were entered.

When the project is uploaded then the attributes are automatically shown .The attributes are assigned to the rules.

Before this process is done the rules for the validation is entered.

The below screen is for creating rules

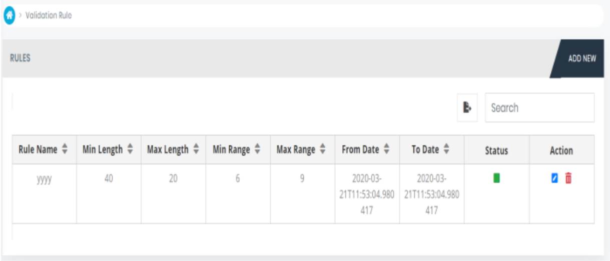


ADD / EDIT RULE

FORMAT mm/dd/yyyy

Rules can be uploaded or created we can upload rule through file. When the rules are created by Admin the user can add rules to their attributes.

After the rules are entered those will be saved and shown like this in the webpage



Rule Name	Min Length	Max Length	Min Range	Max Range	From Date	To Date	Status	Action
yyyy	40	20	6	9	2020-03-21T11:53:04.980417	2020-03-21T11:53:04.980417	🟩	

These details are directly entered into the database and stored in the database.

As a result of applying rules can give the appropriate and structured data.

PROBLEM STATEMENT

Performing web crawling without wrangling might result in data

As listed below:

1. Unstructured Data
2. Irrelevant/Garbage Data
3. Data that does not meet client's requirements.

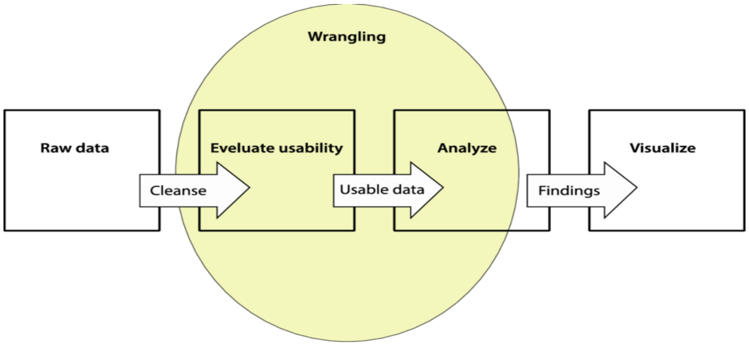
SOLUTION

Data wrangling must be a mandatory process if the data that is pre-processed should result as a meaningful data.

Data wrangling makes sure that the data crawled from the web is

1. Appropriate to the requirements
2. Relevant to the category
3. High value data
4. Structured Data

FLOW DIAGRAM



FUTURE SCOPE

Large Data cannot be validated at once. But in the future that can be validated and will be very useful and easily validated.

TECHNOLOGIES USED

SOFTWARE:

FRONTEND:

1. C#
2. .NET CORE

BACKEND:

PostgreSQL

TOOLS:

1. Microsoft Visual Studio
2. Microsoft Visual Studio Code

CONCLUSIONS

Staying on your path in the forest of information requires a lot of concentration and effort.

However with the help of Data Wrangling process we can validate and give the structured and high valued Data.

When you gain insights and make your business decisions based on them, you gain a competitive advantage over other businesses in your industry. Yet, it doesn't work without doing the homework first and that's why you need data wrangling processes in place.

REFERENCES

[1]https://theappsolutions.com/blog/development/data-wrangling-guide-to-data-preparation/